

ECON M-518: Employee Attrition Analysis

Final Project

Yash Shah
MS Data Science
Indiana University, Bloomington
Indiana, USA
yashah@iu.edu

Ritwik Budhiraja
MS Data Science
Indiana University, Bloomington
Indiana, USA
rbudhira@iu.edu

Abstract — *Employee Attrition (EA) is the gradual reduction in employee numbers. It happens when the size of your workforce diminishes over time – meaning that employees are leaving the companies faster than they were hired. The reasons for attrition are somehow more complicated and differ from organization to organization and country, depending on the work culture in each county. This project talks about a company specializing in the field of computers in the USA and we look at the various factors causing attrition in this company and make substantial predictions about EA in this company in the future. R programming language is used for this analysis.*

Keywords — *Attrition, HR Analytics, EDA, Logistic Regression, Machine Learning Models, R.*

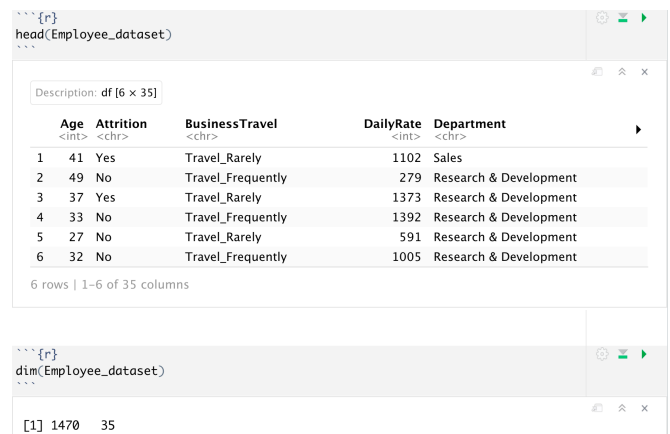
I. INTRODUCTION

In a world where the skill sets required are constantly changing, positions in a company also become obsolete over time. As employees leave and a new future of work emerges, not every role is filled in the same cookie-cutter way. With this, a new world of work means an upgrade of your existing work and being driven by a desire to modernize. This is a way to look at Employee Attrition (EA) in organizations. EA can be problematic as it often reduces talent within the company, but it can be positive by allowing the company to identify and address the issues for its employees. For instance, a high attrition rate could be from employees leaving due to poor work culture. Only by investigating the reasons for this employee attrition can HR make the appropriate changes to improve the organization's work culture for the other employees. While companies usually try to avoid EA, it can sometimes help cut their labour costs and attract new employees with fresher talents. For many, leaving a job is due to personal reasons. However, understanding tendencies and 'leaving patterns' can be crucial for a company to understand. This is often the first step in preventing EA in the future.

II. DATA DESCRIPTION

The dataset used for this project has been taken from the internet. The data published was not authentic and it is generated artificially. The idea for this data was that the company wanted subjects to use this data and perform descriptive and predictive analysis. The data has

1470 rows and 35 columns with variables like age, salary, work-life balance, satisfaction etc. The variables that look the most promising and would be a big contributing factor in this analysis would be — Attrition (Training): Takes 'Yes' or 'No'; Age: Takes a range of numbers from 18 to 60 years; Department: Sales, R&D, or Human Resources; Total Working Years: Number of years an individual has worked for; Monthly Income: Salary in USD; Distance from Home: Distance from home in miles; and Years since last promotion.



```
{r}
head(Employee_dataset)

Description: df [6 x 35]

  Age  Attrition  BusinessTravel  DailyRate  Department
<dbl> <chr>      <chr>           <dbl>    <chr>
1   41      Yes      Travel_Rarely      1102     Sales
2   49      No       Travel_Frequently    279     Research & Development
3   37      Yes      Travel_Rarely      1373     Research & Development
4   33      No       Travel_Frequently    1392     Research & Development
5   27      No       Travel_Rarely       591     Research & Development
6   32      No       Travel_Frequently    1005     Research & Development

6 rows | 1-6 of 35 columns

{r}
dim(Employee_dataset)

[1] 1470 35
```

Figure 1: Structure of the Dataset (head and dimensions)

III. DATA MANAGEMENT IN R

Managing the data for this was not very difficult. By using *glimpse()* we can see every column in the data frame at once and we can have a broad look over the data types, variable types and labels in our dataset as shown in *Figure 2*. Going deeper into the variable types, we encounter numerical and categorical variables. These categorical variables can then be classified as binary (Yes/No, 1/0), ordinal (0-4, 1-5), nominal (Department, BusinessTravel) and so on. As the first step, we change values for Yes/No to 1/0 for two variables 'Attrition' & 'OverTime'. We then remove the variables which would not help our analysis. For instance, the variable 'Over18' has the value 'Yes' for all the observations. Similarly, variables like 'StandardHours' and 'EmployeeCount' have the same values for all the rows, so we remove them from our dataset. Next thing is to look for any missing data. Luckily, the data does not have any missing values.

The biggest challenge we would have to face with this data would be creating relative data tables for certain variables just to create visualizations. This is because the dataset used is an imbalanced one in many ways. For instance, the

variable ‘Department’ comprises three distinct values and there is the highest amount of attrition seen in the Research and Development Dept. (R&D) but since the number of employees in each department is not equal, thus not advocating this inference. To solve this, we look at the proportions and understand that the Sales department sees the highest attrition followed by Human Resources and then R&D. Another thing that we would have to take care of would be the imbalance in the number of observations seeing attrition and the ones not seeing attrition (shown in Figure 3). This would be an important factor to consider while determining the best approach to implement our predictive model.

```

{r}
Employee_dataset %>% glimpse
Rows: 1,470
Columns: 35
$ Age                <int> 41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35, 29, 31,
$ Attrition          <chr> "Yes", "No", "Yes", "No", "No", "No", "No", _
$ BusinessTravel     <chr> "Travel_Rarely", "Travel_Frequently", "Travel_Rare...
$ DailyRate         <int> 1102, 279, 1373, 1392, 591, 1005, 1324, 1358, 216, _
$ Department         <chr> "Sales", "Research & Development", "Research & Dev...
$ DistanceFromHome  <int> 1, 8, 2, 3, 2, 2, 3, 24, 23, 27, 16, 15, 26, 19, 2, _
$ Education          <int> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1, 2, 3, 4, 2, _
$ EducationField     <chr> "Life Sciences", "Life Sciences", "Other", "Life S...
$ EmployeeCount      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, _
$ EmployeeNumber     <int> 1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14, 15, 16, 18, _
$ EnvironmentSatisfaction <int> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3, 1, 4, 1, 2, 3, 2, 1, _
$ Gender            <chr> "Female", "Male", "Male", "Female", "Male", "Male", _
$ HourlyRate        <int> 94, 61, 92, 56, 40, 79, 81, 67, 44, 94, 84, 49, 31, _
$ JobInvolvement     <int> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3, 4, 2, 3, 3, 2, 4, 4, _
$ JobLevel          <int> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1, 3, 1, _
$ JobRole           <chr> "Sales Executive", "Research Scientist", "Laborato...
$ JobSatisfaction    <int> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3, 2, 3, 3, 4, 3, 1, 2, _
$ MaritalStatus      <chr> "Single", "Married", "Single", "Married", "Married...
$ MonthlyIncome      <int> 5993, 5130, 2090, 2909, 3468, 3068, 2670, 2693, 95, _
$ MonthlyRate       <int> 19479, 24907, 2396, 23159, 16632, 11864, 9964, 133, _
$ NumCompaniesWorked <int> 1, 6, 1, 9, 0, 4, 1, 0, 6, 0, 0, 1, 0, 5, 1, 0, _
$ Over18            <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", _
$ OverTime          <chr> "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No", _
$ PercentSalaryHike  <int> 11, 23, 15, 11, 12, 13, 20, 22, 21, 13, 13, 12, 17, _
$ PerformanceRating  <int> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, _
$ RelationshipSatisfaction <int> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2, 3, 4, 4, 3, 2, 3, 4, _
$ StandardHours      <int> 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, _
$ StockOptionLevel  <int> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1, 1, 0, 1, 2, _
$ TotalWorkingYears  <int> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, 10, 5, 3, 6, _
$ TrainingTimesLastYear <int> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3, 5, 3, 1, 2, 4, 1, 5, _
$ WorkLifeBalance    <int> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2, 3, 3, 3, 2, 3, 3, 2, _
$ YearsAtCompany     <int> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, 5, 2, 4, 10, _
$ YearsInCurrentRole  <int> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2, 2, 9, 2, _
$ YearsSinceLastPromotion <int> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4, 1, 0, 8, 0, _
$ YearsWithCurrManager <int> 3, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3, 2, 3, 8, 5, _

```

Figure 2: Glimpse of the data

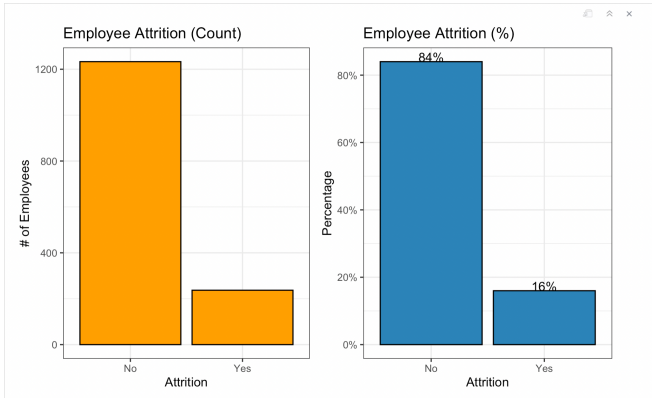


Figure 3: Imbalanced Dataset

IV. RESEARCH QUESTION

There can be many factors upon which we would lay down our research question. The umbrella question would still be to find the causes of Employee Attrition in an organization. This would involve multiple analysis reports on how each variable would be a potential predictor in this project while ‘Attrition’ is our target variable.

V. DATA VISUALIZATIONS & ANALYSIS

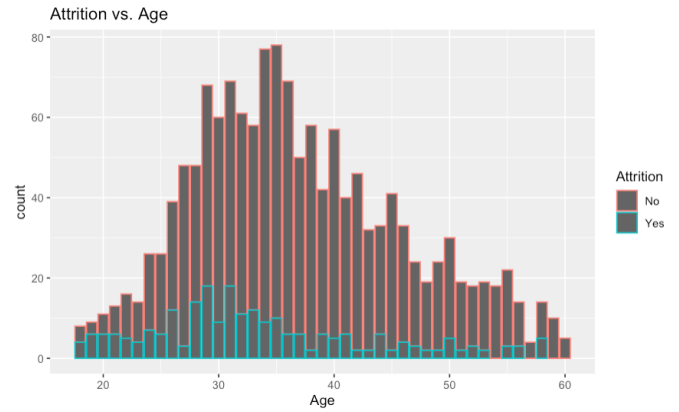


Figure 4: Attrition over Age

As seen in Figure 4, generally, younger people tend to leave the company more. Down the line, people get settled in and are happy with their current situation. Hence, the attrition ratio starts to go down around the age of 30.

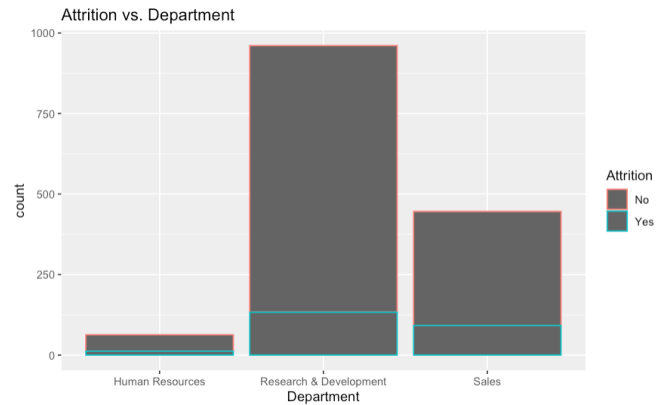


Figure 5: Attrition over Department

As seen in Figure 5, This data comprises of only 3 major departments and looking at the counts and proportions, we can say that the Sales department has the highest attrition (around 25%), followed by the Human Resources department (around 19%). The Research & Development sees the lowest attrition (around 13%).

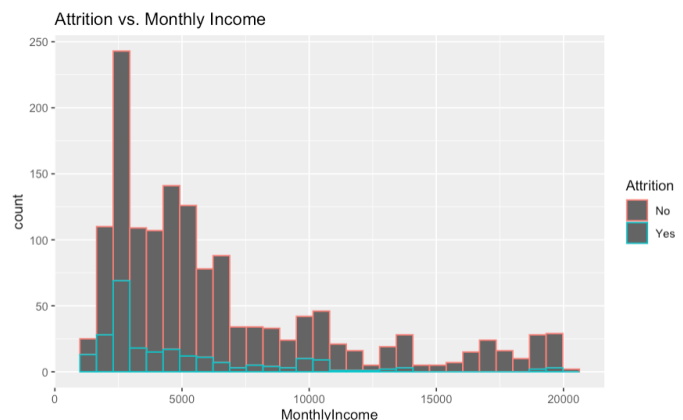


Figure 6: Attrition over Monthly Income

As seen in Figure 6, The attrition rate is evidently high at extremely low income levels—less than 5k monthly—as

shown in the above chart. This falls, but at 10,000, a slight bump is seen that denotes the middle class lifestyle. They frequently change jobs in order to pursue a higher level of living. The flat line (from 14000 - 18000) indicates that there is little risk of an employee leaving the company while the monthly pay is respectable.

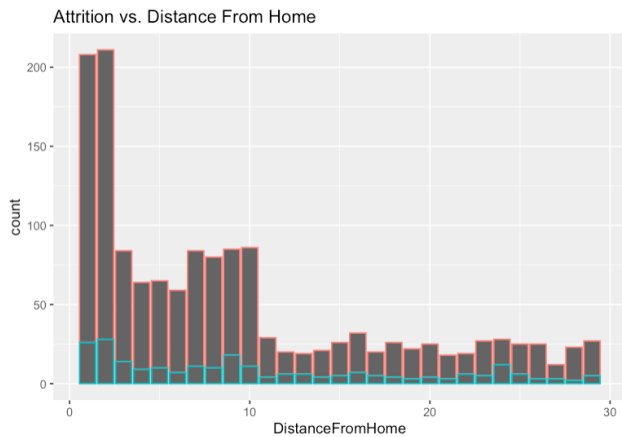


Figure 7: Attrition over Distance from Home

As seen in Figure 7, Most of the people who leave the company are located within than 1-10 miles away from the company. We would expect distance to be a cause of attrition for the employees but it is not so.

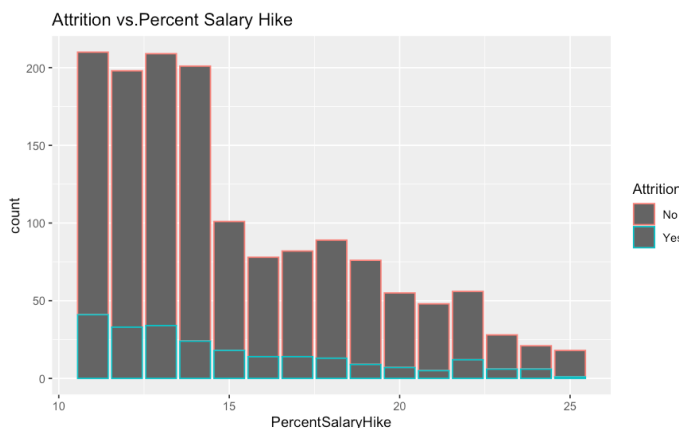


Figure 8: Attrition over Percent Salary Hike

As seen in Figure 7, increased pay encourages employees to perform better and stick with the company. As a result, we can observe that a company where the raise is lesser has far higher risks of losing an employee than one that offers a big raise.

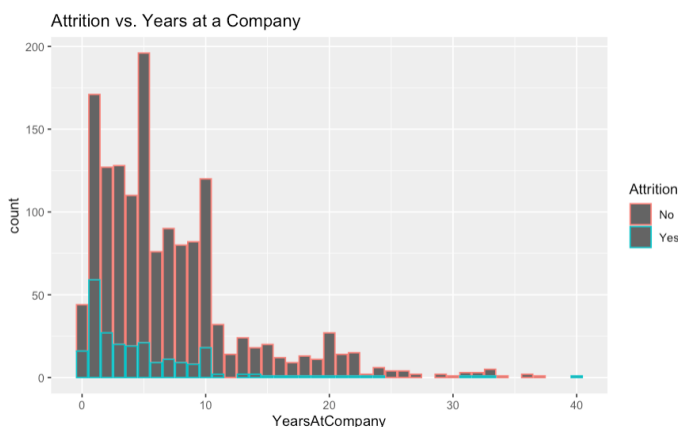


Figure 9: Attrition over Years at a Company

As seen in Figure 9, We can see that the first two bars have a higher ratio of yes to the total count compared to the remaining. This could be probably because of the fact that employees have to stay for the probationary period, and they tend to leave if they are not satisfied with the company after that.

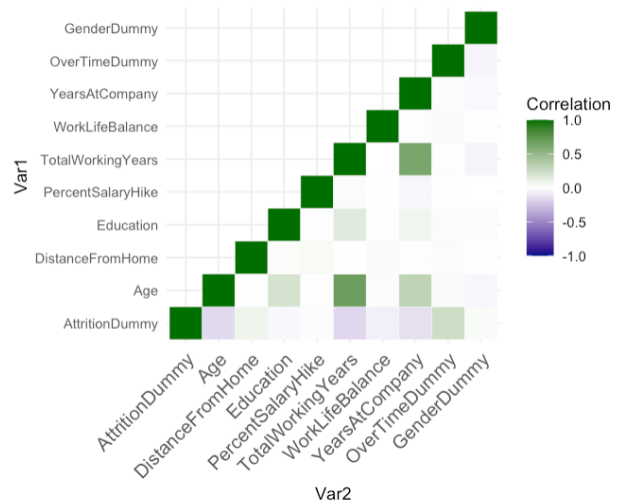


Figure 10: Correlation Heatmap

As seen in Figure 10, On inspecting the visualizations and then correlation matrix heatmap, since we consider Attrition to be our target variable, we can say that attrition is correlated with variables like TotalWorkingYears, YearsAtCompany and Age at the most.

VI. MODELLING & RESULTS

The question of how likely it is for an observation to belong to each group is frequently addressed via logistic regression. This model is frequently employed to forecast the likelihood that an event will occur. Using a logit function, logistic regression's output is transformed as opposed to linear regression's. The result is therefore either 0 or 1. We are interested in forecasting whether an employee will quit (1) or stay (0), hence this approach is advantageous for this issue.

The interpretability of logistic regression is another factor making it the chosen model. The outcome of the response variable (attrition) is predicted by logistic regression using a number of additional explanatory factors, often known as predictors. Our response variable's value falls into one of two categories in relation to this domain: 0 (zero) or 1 (one). The chance of an employee staying with the firm is represented by the number 1 (one), whereas the probability of an employee leaving the company is represented by the value 0 (zero).

On using two models to predict the attrition in employees. Both these models are logit models or logistic regression models use the same target or response variable ie. Attrition and contain different predictors.

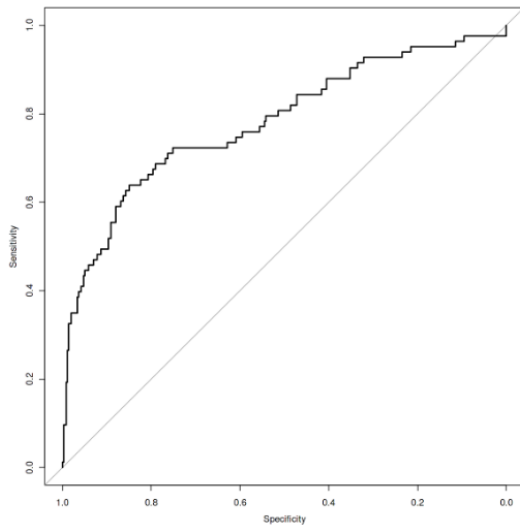


Figure 11: Logistic Regression Model 1

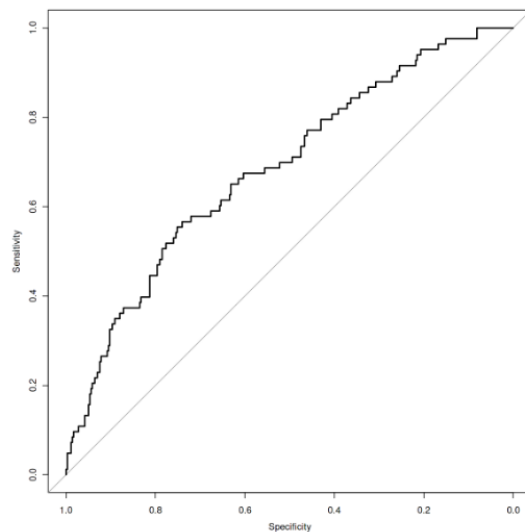


Figure 12: Logistic Regression Model 2

Despite having the word "regression" in its name, Logistic Regression is a type of parametric classification model in the realm of machine learning. Accordingly, logistic regression models produce categorical predictions, such as whether a plant belongs to a particular species or not and have a set number of parameters that rely on the amount of input characteristics.

As shown in *Figures 11 and 12* The Y-axis, as seen, runs from 0 to 1. This is because the sigmoid function always uses these two values as its maximum and minimum, which perfectly satisfies our desire to divide samples into two groups. We may determine the probability that an observation belongs to one of the two categories by computing the sigmoid function of X, which is a weighted sum of the input features, similarly to linear regression. There are a number of methods of evaluating whether a logistic model is a good model. One such way is sensitivity, specificity.

Sensitivity, which is a complement to the false negative rate and is also sometimes referred to as the true positive rate or recall, measures the percentage of actual positives

that are correctly identified as such (for example, the percentage of ill people that are correctly identified as having the condition). Mathematically,

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. Mathematically,

$$\text{Specificity} = \frac{\text{True Negative}}{\text{False Positive} + \text{True Negative}}$$

How do we translate this into language used in business? It is assumed that you are using historical data to execute this test. If a model is effective, you want it to be able to estimate the likelihood that a customer will react to a strategy when in fact the client has already done so. This is an indication of sensitivity. It is probable that a consumer is responding even though the model assumes they haven't, and vice versa. The idea is comparable to a Type 1 or Type 2 statistical error. In a similar vein, if the client has not answered, you want to model as well to present the real picture so that you may base business decisions on the data that are actionable.

For any test, there is always a trade-off between these two terms. Now, one approach to determine whether your model is a good model is to have a high sensitivity and specificity. The accuracy levels are subject to trade-offs, and it relies on the nature of the business. However, in a logistic model, the greater the sensitivity, the better it is relative to other measures.

Logistic Regression Results		
Measures	Model 1	Model 2
Sensitivity	0.8966	0.8633
Specificity	0.9603	0.5861
Accuracy	0.9148	0.8462

Table 1: Results

We can evidently see that Model 1 works better than Model 2 with a higher sensitivity and a higher specificity score. Overall to sum it up, we also look at the Accuracy score and can thus conclude that Model 1 does a better job than Model 2 at predicting attrition in employees.

VII. CONCLUSION & FUTURE SCOPE

Employee Attrition being a very important topic in today's corporate market, it is vital for each company to know the strengths and the weaknesses of their employees and keep the employees happy and satisfied. Apart from the 35 variables that we worked on, there are many more factors that

could cause attrition in employees. We understood that Sales Department has got more Attrition compared to other departments. The employees who are Sales Representatives left the job frequently, and next comes lab technicians. The attrition rate was more for the employees who are single rather than married. Employees who travel more frequently get a higher attrition rate. Employees working more than normal working hours(during overtime) influence the attrition rate. We also understood that logistic regression is a good regression model to work with when predicting binary results in the target variable.

As a part of the future scope, we could use much more complex prediction methods like Decision Trees, which would help us predict the response variable ie. Attrition (Yes/No) and thus decide on each employee's possibility to leave the company. Similarly, we can also use Ensemble methods like Random Forest and Extreme Gradient Boosting which would yield better results.

VIII. ACKNOWLEDGEMENTS

We would like to sincerely thank our instructor and mentor **Prof. Stefan Weiergraeber** for his constant support and constructive criticism which helped us complete this project.

IX. REFERENCES

1. <https://towardsdatascience.com/logistic-regression-explained-from-scratch-visually-mathematically-and-programmatically-eb83520fdf9a/>
2. <https://www.rdocumentation.org/>
3. <https://www.jigsawacademy.com/sensitivity-vs-specificity-in-logistic-regression/>